UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

----------x

UNITED STATES OF AMERICA :

- v. - : 15 Cr. 565 (VEC)

KEVIN JOHNSON, :

DEFENDANT. :

----------x

PLEASE TAKE NOTICE, that the defendant herein, **Kevin Johnson**, will move this Court, before the **Honorable Valerie E. Caproni**, United States District Judge for the Southern District of New York, at a time to be set by the Court, for an order excluding evidence at trial generated by, and testimony about, the Forensic Statistical Tool ("FST"), or in the alternative, granting his request for a hearing pursuant to *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 589 (1993), and granting such further relief as this Court deems just and proper.

DATED: New York, New York
December 1, 2016

Respectfully submitted,
DAVID E. PATTON, ESQ.
Federal Defenders of New York
By: _/s/_____
CHRISTOPHER FLOOD, ESQ.
SYLVIE LEVINE, ESQ.
ROBERT M. BAUM, ESQ.
Attorneys for Defendant
Kevin Johnson
52 Duane Street - 10th Floor
New York, New York 10007

TO: PREET BHARARA, ESQ.
United States Attorney
Southern District of New York
One St. Andrew's Plaza
New York, New York 10007

Attn: **JASON SWERGOLD, ESQ.**
STEPHANIE LAKE, ESQ.
Assistant United States Attorney

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

----------x

UNITED STATES OF AMERICA :

- v. - : 15 Cr. 565 (VEC)

KEVIN JOHNSON, :

DEFENDANT. :

----------x

REVISED MEMORANDUM OF LAW IN SUPPORT OF MOTION TO EXCLUDE DNA-EVIDENCE GENERATED BY THE OCME'S FORENSIC STATISTICAL TOOL (FST)

Federal Defenders of New York
Attorney for Defendant
Kevin Johnson
52 Duane Street, 10th Floor
New York, New York 10007
Tel.: (212) 417-8700

Christopher Flood, Esq.
Sylvie Levine, Esq.
Robert M. Baum, Esq.
Of Counsel

TO: PREET BHARARA, ESQ.
United States Attorney
Southern District of New York
One St. Andrew's Plaza
New York, New York 10007
Attn: **Jason Swergold, Esq.**
Stephanie Lake, Esq.
Assistant United States Attorneys

Table of Contents

Table of Authorities

Cases

Rules

I. Introduction

In the public's name, computers at the New York City Office of the Chief Medical Examiner (OCME) run different programs every day. Some manage payroll, some support the laboratory's many functions or keep the elevators running, while others send and receive email. All of this software does exactly what it is expected to do. Obvious, tangible consequences follow upon failure: rent checks bounce, elevators halt in the shafts, email stops.

The OCME runs one program that differs from the rest: The Forensic Statistical Tool (FST).[1] It purports to calculate a likelihood ratio (LR) representing the probability of a piece of evidence given two hypothetical alternative scenarios. The LR is an inherently unverifiable statistic. How FST claims to calculate the LR may be well-documented[2], but how the program actually operates within its "black box" is a carefully guarded secret.

To our knowledge, the Court's July 18, 2016 Protective Order permitted the first outside review of FST's source code. This review reveals FST to be performing "undocumented behavior."[3] That is, the code performs a routine that is not set forth in its documentation. This function has a marked effect on the LR that clearly disfavors the defendant. Inside its black box, FST is something different than what it has been made out to be.

Implications ring from these facts. It can presently be said with certainty that (1) the OCME never made clear that its flagship, copyrighted software differed in any way from its documentation; (2) the issue is significant as even small changes to a LR are shown to have substantial effects on its product; and (3) the FST authors' omission of this hidden function from any documentation, validation, testimony, or any other opportunity already afforded them to identify and explain it severely

---

[1] While similar to what we understood FST to be, the program under review performs an otherwise undocumented mathematical routine that is not reported anywhere in the extensive public and court record surrounding FST, including representations made in this case.

[2] Tr. May 9, 2016, p. 5, ln. 15-19; p. 42, ln 7-11 (referencing publications and studies in which FST authors describe the program's formula.).

[3] See Declaration of Nathaniel Adams, October 27, 2016 ("Adams Decl."), at section 5.6. Attached at Exhibit C.

damages what credibility they may retain. All demonstrate the inadmissibility of either FST or FSTπ.[4]

❖ ❖ ❖

This past September, the OCME issued an announcement that, by January 1, 2017[5], the lab "plans to entirely cease using … FST on new casework"[6] and will be "retiring" the program[7]. FST's era may finally be ending, but its deathwatch has been active for some time. After a year-long admissibility hearing in *Collins & Peaks*[8] resulted in FST's exclusion, there has been uncertainty about its admissibility among the state courts.[9] The evidentiary record developed in *Collins* reveals a chain of shoddy and rushed validation, unfounded assumptions, and compromised data reaching to its origin. This flawed foundation only begs further questions, and to date, no federal court has admitted FST results over a defense objection.

❖ ❖ ❖

FST does not square with constitutional principles of fairness; the program truly threatens Mr. Johnson's fundamental rights; and it blurs the coherence of federal

---

[4] For purposes of this Motion, where necessary to distinguish the code received pursuant to the July 18, 2016 Protective Order from that described in the public record, we refer to the one received pursuant to the Protective Order as FSTπ.

[5] See Memorandum from Timothy D. Kupferschmid, Chief of Laboratories, Office of Chief Medical Examiner, on Implementing New Technologies (Sept. 19, 2016). Attached at Exhibit K.

[6] The OCME will continue to use FST on samples that it already processed using Identifiler, the kit also scheduled for scuttling on January 1. *Id.* How long of a half-life this represents for FST is a function of how many samples OCME has in the pipeline. Hopes are it is not many.

[7] In conjunction, the OCME is also retiring its current STR kit and analysis software, signaling what must be major validation overhaul. OCME is finally discontinuing LCN testing. *Id.*

[8] *People v. Collins & Peaks*, Ind. Nos. 80077-2010, 7690-2010 (Sup. Ct. Kings Cnty. Nov. 7, 2014) (hereinafter *"Collins"*). Electronic copies of relevant excerpts of *Collins* Transcripts are available to both the Court and opposing counsel ("*Collins* Transcript").

[9] A wide gulf divides the *Collins* decision from that of the only other court to hold an admissibility hearing regarding FST. In *People v. Rodriguez*, Ind. No. 5471-2009, the court determined that FST was generally accepted and thus admissible as it "rests firmly upon two pillars, [PCR-STR] DNA analysis and the likelihood ratio." *Id.* Decision and Order, at p. 8 (Sup.Ct. New York Co. Oct. 24 2013) (Carruthers, J.). The *Rodriguez* court seized on a diversionary issue, as no defendant would challenge the general acceptance of Baysean statistics. The issues with FST are not so perfunctorily dismissed, as the *Collins* court, conducting a hearing the following year, recognized.

criminal procedure. The scope of FST's prejudicial harm is clear: While the prosecution of this case may not turn on DNA, the use of the DNA in this case depends entirely upon FST. Initial tests were not inculpatory of Mr. Johnson. He is not included in any sample. It was not until FST calculations were performed that "very strong support" arrived for the hypothesis that he could be in the samples at all. Thus FST and its underlying methodology figures prominently here.

A. FST's LR cannot help the trier of fact as it isn't derived from a reliable principle or method.

This "support" is more illusion than substance. FST's claim to an eight-digit LR on these facts dangerously implies certainty. A LR may resemble a rarity statistic, but it is not one. No LR formula is straightforward, but FST is particularly confusing, compounding speculation through its algorithm. Further inspection only tends to clarify how confusing and unsuitable this formula is to criminal trials.

The very existence of a hidden function in the source code demonstrates this point. It is not a subtle effect:[10] The function will drop an entire locus from the LR based on frequency data calculated for alleles.[11] Though obviously contrary to FST's documentation, this routine appears to have been operative without being noticed since FST's inception. Only an imaginary number like the LR unbound from any objective standard could go on undetected for so long.

Unearthing the fault has taken the monumental effort of a review of the source code – an effort made considerably more difficult by the OCME's opposition. Absent this review, the fault would have remained undetected, not for lack of its operation, but for lack of any objective grounding of FST's LR. As our review parses through the sprawling code, further questions are just emerging over what

[10] The effect is stark. We have verified an effect in a complex mixture from the FST validation study. Solely accounting for the action of the "undocumented behavior", the LR reduced to half of its original size. As we have urged on this Court before, LR statistics are not verifiable by reference to an objective benchmark. But viewed through the lens the "undocumented behavior" affords – the ability to alter one variable while keeping all others constant – FST is a powerful, highly sensitive, if uniquely unvalidated algorithm.

[11] It is highly objectionable. The other data lost from that locus–or from those loci–may have been working in the equation on behalf of the accused. At best, the authors never considered this possibility. What is certain is that the authors never disclosed the presence of this locus-deleting function to anyone in the public sphere.

notice the OCME provides to those whose formula has been altered from published methods, or whose loci have been discarded by this or some other function. When we sought the code, what informed the lab's certainty that we were misguided? Has FSTπ been disclosed to any person before? Are any OCME personnel themselves even aware of it? Clear answers are not evident.

B. FST is not based on sufficient facts, data, or sound methodology.

The OCME got a lot wrong with FST, but it fundamentally misconstrued what is relevant about probability, statistics, and their relevance in criminal law. These measures are not means to an end, useful simply because they become weaponized in court. Even if it explicitly was the aim of FST's architects to "provide[ ] quantitative weight … to interpretations that would otherwise be qualitative[ ]"[12], interpretations drawn by the lab's dangerously-sensitive Low Copy system, the relevance in statistics is not the answer, the "weight", but it is the question, or proposition. FST's questions are inherently prejudicial, overly simplistic, and simply not constructed to achieve relevant answers.

The program operations are literally festooned with untested assumptions. Even defining features of OCME's process are supported by nothing more than an untested hypothesis. Emblematic is the lab's repackaging of the quantitation, or "quant" estimate – a measure with a staggering 30% error rate – as the lodestar of the LR calculation. The OCME did not test or validate its assumptions about repurposing the quant estimate. Nor does FST in any way account for the yawning error rate at the heart of its calculation.

C. FST was never validated for this purpose.

FST poses as neutral, but actually takes a side. Its biases are deep, mathematical, and monstrously difficult to reveal. It is rigid. Neither itself expert enough to assist, nor flexible enough to be put to effective use by an expert in assisting a factfinder. FST may calculate a number, but its results don't offer a meaningful answer to a relevant question. Saddled with an unsound, unscientific purpose, FST could not escape the flaws of its design. It both over-and under-shoots its target.

Over-reaching, FST embraces too much: Testing the prosecutor's hypothesis against a putative "defense hypothesis", it attempts to solve the case by itself. The program invades both factual and legal domains, instructing which factors and

---

[12] Adele A. Mitchell, *Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop in*, 6 Forensic Sci. Int'l: Genetics 749-761, 760 (2012). ("Mitchell et al.") Attached at Exhibit M.

inquiries are relevant and which are to be ignored. In its framing of the LR, FST implicitly yet fundamentally misstates the law. FST shifts constitutional burdens, relieving the government and encumbering the defense. Requiring Mr. Johnson to articulate a positon on the evidence, the program wrongly assumes what that position must be, states it on his behalf, and adds further injury to constitutional insult by managing to underestimate how the defense position is evaluated. Then, having calculated the likelihood of the case evidence in light of these unfounded assumptions, FST offers its results of a loaded hypothetical contest, scoring the winner and further explaining how well the government did in its government-sponsored test.

But why should the proponent of evidence also be allowed to offer a lopsided statistical proof that falsely claims to appraise the range of possible scenarios with the evidence, while also endorsing its own side with a qualitative grade? This is just what FST does.

At the same time, FST accomplishes far too little. To "explain" evidence FST always assumes allelic drop-out: The program is custom-designed to explain evidence principally in these terms. Yet it does so inaccurately. Drawing too close a focus on one possible variable, drop-out, FST either under-values or altogether misses other important issues. This leads to obvious inaccuracies that FST papers over. Crucial factors that affect frequency of DNA evidence are silenced as the program undervalues or ignores race and ethnicity, family and relatedness, the frequency of observed alleles, the common fact of allele sharing, and the effect of error the process.

FST cannot help the trier of fact to understand evidence. But it was never designed to. Rather, it was created with a specific function and for a specific purpose. Its function as the first and only, and surely the final, actuarial genotyping tool, FST devises a LR with in-house data. It does nothing else. Constricted to this one function, the program is limited to a single purpose: "to provide[ ] a quantitative weight to interpretations that would otherwise be qualitative or where no conclusions could previously be drawn."[13] This rationale does not seek to offer a neutral explanation for evidence, but to enhance weak evidence. It is based on a fundamental misapprehension of the value of statistics to a trier of fact. It also indicates what will be shown below, an extensive record of biased, unsound, unscientific, and unreliable methodologies strung together with the purpose of

[13] *Id.*

creating a program that produces a LR drastically favoring the prosecution – the clearest example of this perhaps being the hidden function of the "undocumented behavior." Repackaging this history into some analogue of a scientific validity now can only fail, as a program so devised cannot fairly assist the trier of fact in a federal criminal trial.

❖ ❖ ❖

There is a serious concern that the sheer number of legitimate objections, the depth we believe required to explain them, the necessity of being thorough, and the Court's admonition that this submission be comprehensive will collude and overwhelm its argument in details. This concern is central to the very purpose of this motion. Confronting the network of speculation, imaginary numbers, and unfair prejudice that combine in a FST report while before a jury would impose obstacles to fairness, obstacles that should only exist where the evidence can clearly be said to comport with scientific standards. The fact is, at every stage, FST fails to meet minimal measures of reliable methodology. Whether it be the program's definition or construction, its testing or validation, or its implementation or application here, the OCME did not devise a formula that could be fair, reliable or helpful in a federal criminal trial. Moreover, FSTπ's application here has been wholly unexplained. It should be excluded.

II. Procedural History

Mr. Johnson is charged in this case with one count of possession of a firearm after having been previously convicted of a felony.

Shortly before midnight on April 12, 2015, police were called to 3035 White Plains Road in Bronx County. The officers conducted a search of Apartment 5D. The circumstances regarding that search were examined at a hearing before the Court on December 4, 2015, where officers testified that inside the apartment, they recovered two firearms. At the time they were recovered, each firearm was inside of a separate white sock. Both socks, with the firearms within, were retrieved from behind a refrigerator. The motion to suppress was denied.

Mr. Johnson was initially arrested for this charge directly thereafter and taken into custody on April 13, 2015, at approximately 1:00 A.M. On April 16, 2015, a criminal complaint was filed in the Southern District of New York. The Government presented the case to a grand jury in this District and the grand jury declined

to return an indictment. Therefore, the Government moved to dismiss the complaint. At the time of the April 2015 grand jury presentation, the Government was not in possession of any DNA results.

On August 19, 2015, the Government returned to a grand jury in this District. Included in the evidence presented to the second grand jury were the FST-based DNA results that the Government had obtained in the interim. The grand jury returned the instant indictment.[14] Mr. Johnson was arraigned and entered a plea of 'not guilty' before the Honorable Valerie E. Caproni on September 1, 2015.

III. DNA testing in this case

On April 13, 2015, the Bronx Evidence Collection Team (ECT) of the NYPD responded to the 47th Precinct regarding the two recovered firearms.[15] The ECT swabbed both firearms and submitted six DNA swabs to the NYPD laboratory.[16] The six swabs were from:

- Trigger/trigger guard
- Front/back strap and grips
- Slide grip grooves and mag release
- Trigger/trigger guard #2
- Front/back strap and grips #2
- Cylinder/cylinder release and hammer #2.[17]

According to a report dated May 26, 2015, the OCME determined that, regarding four of the samples, human DNA was detected but was insufficient for STR DNA typing: trigger/trigger guard, front/back strap and grip, trigger/trigger guard #2, cylinder/cylinder release and hammer #2.[18] Regarding the remaining two samples, a sample was extracted from the "slide grip grooves and mag release," but no further typing or testing was performed.[19] On the "front/back strap and grips #2," the

---

[14] *See* Indictment, 15 Cr. 565 (VEC).

[15] Voucher no. 2000419553, NYPD Property Clerk Invoice, April 13, 2015. Attached at Exhibit V.

[16] *Id.*

[17] *Id.*

[18] Laboratory Report, OCME, May 26, 2015. Attached at Exhibit D.

[19] *Id.*

OCME conducted STR DNA typing using an amplification kit and "a mixture of DNA was found."[20] The OCME determined: "The DNA profiles of the individual contributors to the mixture could not be determined; however, the results are suitable for comparison."[21] All six swabs were consumed during their respective testing procedures.[22]

According to a report dated July 17, 2015, the OCME conducted STR DNA typing using an amplification kit on the "slide grip grooves and mag release" (the swab they did not test on May 26) and "a mixture of DNA was found."[23] The OCME determined: "The DNA profiles of the individual contributors to the mixture could not be determined; however, the results are suitable for comparison."[24]

According to a report dated August 5, 2015, the OCME set out to compare Kevin Johnson's DNA to those in the samples of the "slide grip grooves and mag release" and "front/back strap and grips #2."[25] First, the lab created a profile for Mr. Johnson using STR DNA typing. Second, the OCME compared his profile to the two samples and "likelihood ratios were calculated."[26] In this context, likelihood ratios are calculated by the OCME using FST.

After calculating a likelihood ratio, the OCME assigns a qualitative label to the ratio, according to the following chart:

The OCME's comparison yielded the following claims:

Reported value	Qualitative interpretation
1	No conclusions
1 to 10	Limited support
10 to 100	Moderate support
100 to 1000	Strong support
Greater than 1000	Very strong support

Excerpted from OCME Letter, August 5, 2015.

---

[20] *Id.*

[21] *Id.*

[22] *Id.*

[23] Laboratory Report, OCME, July 17, 2015. Attached at Exhibit E.

[24] *Id.*

[25] Laboratory Report, OCME, August 5, 2015. Attached at Exhibit F.

[26] *Id.*

"The DNA mixture found on swab of 'slide grip grooves and mag release' [hereinafter "slide"] is approximately 156 times more probable if the sample originated from Kevin Johnson and one unknown, unrelated person than if it originated from two unknown, unrelated persons. Therefore, there is strong support that Kevin Johnson and one unknown, unrelated person contributed to this mixture, rather than two unknown, unrelated persons."[27]

"The DNA mixture found on swab of 'front/back strap and grips #2' [hereinafter "strap"] is approximately 66 million times more probable if the sample originated from Kevin Johnson and two unknown, unrelated persons than if it originated from three unknown, unrelated persons. Therefore, there is very strong support that Kevin Johnson and two unknown, unrelated person contributed to this mixture, rather than three unknown, unrelated persons."[28]

When conducting FST analysis, the OCME generates Comparison Reports. Those reports indicate that the DNA template amount for the slide was 119 pg[29], and for the strap was 175 pg[30].

IV. Struggle for the source code

Given the use of a "black box" program to implicate Mr. Johnson – a technology never before introduced in federal court – the defense sought to evaluate the FST software, or source code. After that initial request in court, the parties adjourned to ascertain whether the OCME – a public laboratory – would turn it over voluntarily. They refused to do so.[31]

By letter dated May 4, 2016, Mr. Johnson requested that the Court issue a subpoena for the source code, pursuant to Rule 17 of the Federal Rules of Criminal Procedure. A series of subsequent letter briefs set out the parties' positions on the merits under the standard discussed in *United States v. Nixon*, 418 U.S. 683, 700 (1974) and its progeny. On June 7, 2016, by written Order, the Court granted Mr. Johnson's request for a subpoena under Rule 17(c). The OCME then sought to oppose the disclosure of the source code via several letters which were ultimately treated as motions to quash and denied. The Court ordered that Mr. Johnson's

---

[27] *Id.*

[28] *Id.*

[29] Forensic Statistic Comparison Report, slide, July 21, 2015. Attached at Exhibit P.

[30] Forensic Statistic Comparison Report, strap, July 21, 2015. Attached at Exhibit Q.

[31] *See* Letter from Government, 5/3/16, at Dkt. No. 37.

counsel and the OCME negotiate the terms of a protective order. A joint proposal was submitted to, and approved by, the Court. Our experts promptly began a review of the code. To our understanding, this marked the first time anyone outside of the OCME has inspected FST's true functions.

V. FST exemplifies problems with complex mixture analysis

No longer is the question in DNA analysis *Is there a match*, *yes or no*? Turning from basic matching of single-source samples to analyzing complex mixtures introduces entirely different interpretative and technological choices. Complex mixtures (those with more than two contributors) are "biological samples from multiple unknown individuals in unknown proportions."[32] These commonly occur in evidence samples. Swabs of surfaces can recover tiny quantities of biological material left by multiple individuals (e.g. from a doorknob, or a gun).[33] Such complex mixtures are far more complicated to analyze, and it is "inherently difficult and even more so for small amounts of DNA".[34] It is in the analysis where the DNA's power to discriminate breaks down. Subjective assessments that may have a marginal effect in single-source casework are central to complex mixture cases analysis. As a presidential commission recently found, "[t]he fundamental difference between DNA analysis of complex-mixture samples and DNA analysis of single-source and simple mixtures lies not in the laboratory processing, but in the interpretation of the resulting DNA profile."[35]

Interpretation of mixed profiles is more complicated because "each individual may contribute two, one or zero alleles at each locus; the alleles may overlap with one another; the peak heights may differ considerably, owing to differences in the amount and state of preservation of the DNA from each source; and the 'stutter peaks' that surround alleles can obscure alleles that are present or suggest alleles that are not present."[36] As a result, "It is often impossible to tell with certainty which alleles are present in the mixture or how many separate individuals contributed to the mixture, let alone accurately to infer the DNA profile of each

---

[32] Section 5.2, Executive Office of the President's Council of Advisors on Science and Technology ("PCAST") at 75. Attached at Exhibit L.

[33] *Id.*

[34] *Id.*

[35] *Id.*

[36] *Id.* at 75-76.

individual."[37] "Instead, examiners must ask: 'Could a suspect's DNA profile be present within the mixture profile? And, what is the probability that such an observation might occur by chance?'"[38]

Without clear guidance from the evidence itself, as with a single-source sample, there is risk that examiners' expectations about the evidence influence the interpretation. "[A]t least in complex situations (such as with DNA mixtures) DNA does require and rely on human examiners making a variety of subjective judgements that are susceptible to bias."[39] Bias affects the accurate interpretation of even relatively simple mixed samples. When more challenging samples are present, as in Kevin Johnson's case, the effect can be dangerous. This is a systemic problem. When analysts were found to have misapplied the CPI statistic at a lab in Texas, 50,000 cases required reexaminations.[40]

The Executive Office of the President's Council of Advisors on Science and Technology (PCAST) appropriately orients the issue, placing the problems with mixture interpretation and examiner bias at the center of the discussion. According to the PCAST report, the future appears to lie in probabilistic genotyping programs that attempt to remove human subjectivity from the interpretation of complex mixtures, observing that such efforts "clearly represent a major improvement over purely subjective interpretation."[41] The report does not, however, endorse the admission of probabilistic genotyping methods.

The PCAST report cautions that, before using a method, the limitations of a proposed method should be scrutinized to determine four key areas of performance:

---

[37] *Id.* at 76.

[38] *Id.*

[39] *See* Itiel E. Dror & Greg Hampikian. *Subjectivity and bias in forensic DNA mixture interpretation*, 51 Science & Justice 204-8 (2011). Dror and Hampikian detail a case where the DNA mixture was evaluated by analysts who knew that the evidence was from a rape case in a jurisdiction where corroboration was required and found the defendant 'could not be excluded; the same sample was sent for evaluation by 17 other DNA examiners who were provided with no information. The authors discovered that only one of the unbiased reviewers came up with the same sample; 16 had an alternative opinion. Of those 16, 14 concluded the defendant was 'excluded' and four said the results were 'inconclusive'. The variety alone supports the presence and impact of subjectivity. Attached at Exhibit R.

[40] PCAST at 77-78.

[41] PCAST at 79.

(1) How well does method perform as a function of the number of contributors to the mixture? How well does it perform when the number of contributors to the mixture is unknown?
(2) How does the method perform as a function of the number of alleles shared among individuals in the mixture? Relatedly, how does it perform when the mixtures include related individuals?
(3) How well does the method perform – and how does accuracy degrade – as a function of the absolute and relative amounts of DNA from various contributors? For example, it can be difficult to determine whether a small peak in the mixture profile represents a true allele from a minor contributor or a stutter peak from a nearby allele from a different contributor. …
(4) Under what circumstances – and why – does the method produce results (random inclusion probabilities) that differ substantially from those produced by other methods?[42]

In answering to these requirements, FST fails to meet even baseline expectations. Not designed with discipline or subjected to truly rigorous testing or validation, FST does not withstand scrutiny now.

FST does not sufficiently account for interpreter bias because the OCME protocols do not sufficiently account for the bias:

> The purpose of these guidelines is to provide a framework which can be applied to the interpretation of STR results in casework. … [N]ot every situation can be covered by a pre-set rule. Equipped with these guidelines, *analysts should rely on professional judgment and expertise*.[43]

The OCME analyst decides how many contributors are in a sample, and whether the contributor is considered "major" or "minor." The protocols themselves acknowledge this subjectivity: "If the analyst cannot decide between two and three contributors after applying the above guidelines…the analyst's discretion

---

[42] *Id.* at 79-80.

[43] *See* NYC Office of Chief Medical Examiner, Forensic Biology Protocols for Forensic STR Analysis, April 15, 2016 (working version) (available at: http://www.nyc.gov/html/ocme/downloads/pdf/Fbio/Protocols%20for%20Forensic%20Mitochondrial%20DNA%20Analysis.pdf) at STR Results Interpretation, Section VII (emphasis added) ("OCME Protocols").

should be used when doing this determination."[44] These choices are made *before* FST is run.[45]

Thus, the very inclusion of FST amidst the PCAST category of "objective methods" and "probabilistic genotyping" programs is misplaced. Employed only *after* the examiner interprets the sample, the program amplifies rather than replaces biases inherent in the analysis.

VI. Likelihood Ratios in this context stand to confuse the jury

Likelihood ratios compare the odds of two conjectural propositions. They are common in areas plainly inapposite to this case such as paternity testing, medical diagnoses, and horse betting. In federal criminal trials, however, with one recent exception,[46] to our knowledge, no LR has ever been admitted over objection. There is sound justification for this bar: The comparison of two hypothetical scenarios is an exercise in pure conjecture.

As presented by the OCME, the LR equation seems purposefully designed to be inadmissible. With a dangerously misleading legal formulation, the OCME pits a "prosecutor's hypothesis" that here counterfactually presumes the presence of Mr. Johnson's DNA in each mixed sample, against a "defendant's hypothesis" that, while attributed to Mr. Johnson, was never chosen by him. Rigidly framing the subject as a straightforward "either/or" proposition, the equation mischaracterizes the government's task to prove each element beyond a reasonable doubt, lightening it to something akin to a civil preponderance standard. Further, the obligatory "defendant's hypothesis" imposes a burden on Mr. Johnson's right to remain silent,

---

[44] *Id.*

[45] Using the allele count alone, as FST does, "there is a risk of understating that number." Michael D. Coble et al. *Uncertainty in the number of contributors in the proposed new CODIS set*, 19 Forensic Sci. Int'l: Genetics 207-211 (2015) (the additional number of loci – i.e. more than 13 – in the new kit configurations reduces the probability of incorrectly identifying the true number of contributors compared with the existing set of 13). Attached at Exhibit T.

[46] In *United States v. Pettway,* 2016 WL 6134493, 12 Cr. 103 (W.D.N.Y. October 21, 2016), the defense moved in an 8-page, double-spaced brief to exclude the results of STRmix. The effort was not supported by affidavit, declaration, or otherwise by exhibit of any other kind. Unsurprisingly, the motion was denied on the papers on 10/21/16.

suggesting that he must have a defense to present, deciding what it should be, then stating it for him.[47]

Exacerbating these foundational problems, FST calculates the LR, a statistic that purports to "explain the evidence"[48], but which is neither intuitive nor understood. Distinctively jutting into the realm reserved for the factfinder, FST numerically evaluates each side's "hypothesis", scoring the winner. Nowhere else in criminal law is the ultimate issue so quantified.

Beyond these grave concerns, the unscaled numerical product of FST's LR routine[49] is modified by a conditional statement. This essential if highly objectionable component of FST's design determines when a LR offers "support", "strong support", or "very strong" support for the prosecutor's hypothesis. The lines between each category are arbitrarily drawn, and the conclusions instruct the jury how to interpret what is otherwise impossible to understand, precisely what Federal Rule of Evidence 704 prohibits.

VII. FST is flawed in its design, validation, and implementation

The design flaws, assumptions, shortcuts, validation errors, and undocumented procedures by the OCME that are detailed in this section demonstrate the unsuitability of presenting the LR-based results of the program to a jury.

1. Pre-set drop rates ignore case-by-case variables

The authors of FST repeatedly opted for pre-set one-size-fits-all formulas to apply to samples across the board, rather than allowing the sample itself to dictate the variables at issue. For example, the OCME chose an arbitrarily limited number of

---

[47] As discussed below, for a given three-person mixed sample, it is flatly insufficient to impose upon the defense a theory that the defendant is not present and "three random people" are present in the mixed sample. Questions of the true number of contributors to the mixture, the ratio of their presence in the sample, the relatedness of those contributors, the relative age of the DNA in the mixture, and the accurate statistical weighting of the observed alleles all go to what might properly called a "defense hypothesis."

[48] Mitchell et al. at 752.

[49] Notably meaningless without context, there are alternatives for objectively framing the LR that the OCME has undertaken when concerned about this issue. *See* Mitchell et al.

mixture ratios, despite the innumerable ways a real-world sample can be combined.

Moreover, the OCME set the uniform rates based on samples that were unrealistic and deviated from actual casework. OCME used "mixtures of pristine, good quality DNA extracts."[50] The reality of real world samples makes them inherently distinct from the pristine buccal swabs used by the OCME during the validation of FST. In the real world, degradation, peaks below threshold, the quality of touch-DNA, etc. creates real variation.[51] On a case-specific basis, each one of these phenomena would affect the drop rates – and in turn, the LRs.

2. Quant is an insufficient measure for drop rate

A primary unfounded assumption at the heart of FST's model is how the program assigns universal drop rates according to the amount of template DNA that is measured *prior* to amplification. At each locus, the OCME specifies the probability of drop-in and drop-out for mixtures of across a fixed range of DNA quantity. They are the only lab to make this association, and certainly the only one to base the LR on it.

The OCME further assumed that the proportion of each contributor's DNA is present in the mixture such that they all correspond with a preset ratios and such that a universal drop rate is relevant. There is practically no meaning to this association.

According to Dr. Bruce Budowle, who at the time of *Collins* was a professor at the University of North Texas Health Science Center, and was a senior scientist and lab-head with the FBI for more than twenty five years said:

> Q. In your opinion, is the FST a novel approach?
>
> A. I think it is in the way it's being used because it makes certain assumptions about DNA typing that no one else would do even in standard DNA typing. The main assumption being made is that all the rates for drop-in, drop-out are based on the amount of DNA.[52]

---

[50] Dr. Eli Shapiro, Report on FST Validation ("Shapiro Report") at 2. Attached at Exhibit A.

[51] *Id.*

[52] Budowle testimony, *Collins* Transcript, 12/9/2013, at 793.

A potential correlation between quant and drop out (low drop out when samples are robust and vice-versa), does not give rise to the OCME's conclusion that a universal rate applies to all permutations in a mixed profile.[53]

Even if it were theoretically reliable to use a pre-determined statistic based on overall DNA quantity, the error rate of the quant has an unknown effect on drop-out and LR. Dr. Shapiro explains that because the "quantitation is not accurate, each of these extracts was quantified 9 separate times to calculate an average value before using them to create the artificial mixtures used to determine drop-out rates . . . in casework application, the sample is quantified only once, and is only an estimate."[54] In fact, the true rate of error could be double, or up to 60 percent. The OCME knew this would carry forward to the LR.[55] The team did not preserve the data from any testing on this fundamental assumption, nor did they test any possible exceptions to this rule.[56]

3. OCME failed to take allele masking into account

The lab failed to account for allele stacking in the mixtures they created. Stacking, or masking of alleles is an obvious issue challenging the accurate interpretation of any mixture. "When performing validation studies, it is best to utilize samples that are heterozygous at all tested loci if possible so that allele drop-out can be monitored."[57] This simple concept was ignored by the OCME.

Problems with the study began with its inadequate size. Sampling from far too few donors who all worked at the lab, the OCME didn't even try to assemble an array of profiles that adequately modeled the range of variance necessary to test FST's core assumptions. This fatally skewed drop model data. Nothing in FST's formula corrects for this effect.

---

[53] The inferential leap here is akin to the following claim: Because very tall men need leg room, the purchase of large SUV's correlates with height. Then proceeding to use this hypothesis to devise a model of how often vehicle occupants stop for coffee on the turnpike. It may be true that there is some loose correlation, however unreliable, with the data in the first proposition but the second step is putting the theory to an unsound use.

[54] Shapiro Report relying on Mitchell et al.

[55] Mitchell testimony, *Collins* Transcript, 5/1/13 at 97-98.

[56] It bears observation that the error rate is highly significant as both samples in Mr. Johnson's case are quantified at very low template amounts, 119pg and 175pg.

[57] J. Butler, FUNDAMENTALS OF FORENSIC DNA TYPING, 334 (2011).

Design-level failure has consequences. "First, the drop-out rate will be underestimated []. Second, there will be patterns of allele sharing that recur over and over, and arbitrarily affect loci with more allele sharing vs. those with less allele sharing."[58] Dr. Shapiro concludes: "The poor design of this experiment means that arbitrary and incorrect drop-out rates are used by FST."[59]

4. Drop-out was not "empirically determined"

The OCME's claim that FST is empirical is as misleading as it is routinely made.[60] A close review of the drop study demonstrates the point. Dr. Shapiro's assessment is unsparing. "Actually, none of the drop-out rates used by FST matches the empirical values the OCME obtained in the validation."[61] It was not a localized tweak. "This type of arbitrary adjustment of drop-out values occurred for all mixture types and quantification values.[62]" The differential effect is real.[63]

All drop-out rates – including the arbitrarily imposed ones – were reduced by an unconventional factor: One standard deviation from the 'observed' average rate, which allowed a gaping 66 percent confidence interval. Dr. Mitchell's choice to use one standard deviation instead of two, and her decision to go below instead of above the mean is unpublished. She has testified that the data supporting this choice was not compiled or turned over to the Subcommittee but that it was based on an approximate 20 samples.[64]

The lab calls its wholesale revisions "conservative," which makes little sense.[65] Untested changes to the drop study have had unpredictable effects. The OCME

---

[58] Shapiro Report, relying on Mitchell et al.

[59] Shapiro Report at p. 6.

[60] Mitchell et al. at 750. ("The first step of this development was to empirically determine dropout rates for each locus, genotype and DNA template-quantity, and drop-in rates for HT-DNA and LT-DNA amplification conditions."); accord "Empirically estimated drop-out and drop-in rates were incorporated into likelihood ratio frameworks including the appropriate number of contributors (one to three)." *Id.*

[61] Shapiro Report at 3.

[62] *Id.* at 4.

[63] *See id.* at 3 (Finding OCME-altered rate derived a LR of 185, while the same set with a true empirically-determined rate produced a LR of 73.)

[64] Mitchell testimony, *Collins* Transcript, 5/1/13 at 137-9; 5/2/13 at 17.

[65] Shapiro Report at 3.

does not know how badly their choices have prejudiced the defense, principally because they never bothered to conduct any testing.[66] This is cause for study, not an excuse for carelessness.

5. OCME's definition of "conservative" is inaccurate and misleading

In the Forensic DNA community, the term "conservative" is universally defined as "favoring the defendant"[67]. "A conservative estimate is deliberately chose to be more favorable to the defendant than the best (unbiased) estimate would be."[68]

FST's designers use the term "conservative" in ways that contradict its universally-adopted meaning, that contradict their own results, and that contradict their own prior use of the term. They use "conservative" to describe methods that would result in both *higher* drop-out rates (*i.e.* subtracting one standard deviation instead of two) and *lower* drop-out rates (*i.e.* using the most sensitive instruments for the drop study).[69]

This dissembling goes on in the context of explaining a LR logarithm where changing one variable, even in a manner intended to produce a conventionally conservative effect, can lead to the opposite outcome.

6. Favoring non-contributors over the defendant in the LR

FST's principal architect conceded, however, that as FST is formulated, the intended meaning of the term, and critically, the recipient of the statistical benefit, is not the defendant, but the hypothetical non-contributor – who would be present in both the numerator and the denominator of the LR.

> Q: Now, is there any way for you to know that the lower rate is more accurate as opposed to the higher rate being more accurate?
>
> A: We wanted the more conservative rate.
>
> Q: To you always underestimating drop-out is more conservative?
>
> A: Yes.

---

[66] *Id.*

[67] National Research Council, THE EVALUATION OF FORENSIC DNA EVIDENCE, 214 (1996).

[68] *Id.*

[69] *See* Mitchell et al. at 750.

> Q: Always?
>
> A: Not that a lower drop-out rate will always give a lower LR, that's not what I mean by conservative.
>
> Q: It depends on the sample, right?
>
> A: It depends on – We want a low drop-out rate for none [sic] contributors so by choosing a lower – we want a low LR for none [sic] contributors so by choosing a lower drop-out rate we are erring on the side of lower LR for a none [sic] contributor.[70]

By switching definitions in FST, Dr. Mitchell departed from a basic approach where the accused is afforded every reasonable estimation of the weight of the evidence, to one where there are fewer random matches to the evidence. This approach will inherently strengthen the prosecution's case, at the expense of the defense case. This arbitrary changing of the definition of "conservative" is the operative definition of bias.

Traditional concerns the scientific community has had for not overstating the strength of the evidence against the defendant go back to the earliest days of DNA in the courts. From the National Resource Counsel's first and second recommendations through the OPSAC reviews and the PCAST report, the importance, and the actual meaning, of conservativism has held. Only the OCME, and only FST challenges this order.

7. Failure to test

Putting FST into operation without even a minimal case study, without case-by-case performance testing, or without false-positive testing is neither scientifically valid nor defensible. By its nature, the likelihood ratio (LR) encompasses a range of uncertainty. It is not a precise measure by any means. Rather, extensive research continues to expand and refine the limits and reach of LR calculations in the forensic DNA field. Yet, without testing, the OCME put FST into casework even for low LRs, despite the risk of false-positive results.

Although perfectly consistent with the lab's cavalier approach to the LR, road-testing in real cases is not a scientific method of validating a program. Not only does it expose real people like Mr. Johnson to the risk of false positive association,

---

[70] Mitchell testimony, *Collins* Transcript, 5/1/13 at 113-14.

but because the forays are not controlled they therefore cannot even be called experimental. Rather, the phenomenon is best described as OCME's willingness to attempt usage until failure in real-world scenarios. This is not a valid scientific principle, or a methodology likely to produce reliable results.

8. **The OCME's use of the same drop-out rate for the numerator and denominator is not generally accepted and disfavors the defense**

FST holds the drop-out rate constant between the prosecutor's hypothesis and the defense hypothesis. In fact, choosing a rate is, itself, part of each side's hypothesis. The defense might well have a different theory about the number of contributors, level of degradation, and drop out. It is in the prosecutor's interest to keep the drop-out rate low in the numerator, to minimize the number of potential contributors that could fortuitously be included in the mixture; it is in the defendant's interest to keep the drop-out rate high, to include as many potential contributors that the defendant can point her finger at as the true contributor. The LR is highly sensitive to altered variables, and the parties' interests diverge. FST nonetheless mirrors the variables in the hypotheses it calculates, and cannot juxtapose different variables, regarding for example the number of contributors. For FST to function, both the prosecution and "defense" hypotheses must construe the same number of contributors to the mixture – despite the fact that their interests sharply diverge. Moreover, the OCME analyst chooses the hypothesis that "fits", doubtlessly serving the prosecutor's hypothesis.

But it is a rare defense lawyer who lacks the creativity to see beyond the prosecution's own theory of the facts, and true defense work is far more than simply negating the prosecution's statement with "not-*x*." FST affords none of this defense vibrancy. Even if it did, it cannot calculate a LR juxtaposing a true defense hypothesis against the true prosecutor's hypothesis. It offers little more than a sideshow – although a deeply unfair, highly prejudicial, distracting and irrelevant sideshow, it is a sideshow just the same.

This rigidity is unnecessary and prejudicial. Other scientists have demonstrated that varying the number of contributors between the numerator and denominator can have a conservative effect on the LR.[71]

9. **FST ignores the possibility of relatedness**

---

[71] P. Gill & H. Haned, *A new methodological framework to interpret complex DNA profiles using likelihood ratios*, 7 Forensic Sci. Int'l: Genetics 251-263 (2013). Attached at Exhibit N.

The bluntness of the FST instrument is exacerbated by its inflexibility: its formula ignores the possibility of relatedness among the contributors to a DNA mixture or to any possible suspect. Both the numerator and denominator of the FST likelihood ratio presume that the contributors are "random, unrelated individuals," regardless of the actual or suspected relatedness among the contributors to the actual sample. This is simply unrealistic. The program forces the defense to adopt a highly unimaginative and narrow defense theory that constrains the exercise of constitutional rights. The failure to account for such obvious case-specific factual variables as relatedness[72] in the legal context renders FST unreliable and functionally useless, and un-validated for the purpose it is being employed.

The presence of related individuals in any given complex DNA mixture cannot be predicted.[73] Yet the OCME went on to create a mathematical model that fails to allow for this very possibility. The disparity between the OCME's validation study and the real world samples is striking: Related individuals frequently have access to, touch, mix, or share items. This limitation of FST is not only a shortcoming, it renders it inapplicable to real-world case work.

10. FST ignores peaks below threshold

FST does not consider peaks below the 75 RFU threshold. Peaks in a low-quality sample that do not reach the 75 RFU threshold are systematically excluded, whether they are visibly present or not. Their absence may simply be a matter of peak imbalance, yet they are completely excluded. While that approach may benefit a defendant whose alleles are present below threshold, it will prejudice a defendant who is a non-contributor. It also fails to distinguish between those for whom there is no evidence of allelic activity below the threshold whatsoever from those whose alleles *are* present but just below cutoff. It is not a meaningful distinction, grouping the innocent in with the lucky.

11. Linear interpolation does not work

---

[72] Modeling that provides for either related or unrelated contributors is mathematically possible, and indeed present in other LR models, such as LikeLTD. *See* Dr. Rori Rohlfs testimony, *Collins* Transcript, 11/19/13, at 166-170.

[73] Indeed, the OCME's own FST false positive study included a three-person mixture, selected at random from morgue profiles, which contained so much allele sharing that the authors hypothesized that two of the contributors were brothers.

The revealing flaw in OCME's vision for FST was its assumption that the drop-out rates of pre-quantified DNA amounts would, once tabulated, plot according to a linear distribution. Tracking six different quantities (25, 50, 100, 150, 250, and 500 pg, respectively), the design team expected that the six measured quantities would neatly conform. It didn't work out.

Dr. Budowle and others pointed out that the problem is with the idea, not the data. The linear approach is too simplistic to deal with the vagaries of complex degraded mixtures. Logistical regression analysis – a far more complicated approach to the issue – has been adopted for use in other drop models, particularly Balding's LikeLTD, Loehmueller's LabRetriever, and Buckleton's STRMix. OCME's "retirement" of FST and implementation of STRmix should be viewed in this context.

12. Serial dilution and splitting samples into replicates skews the LR in favor of the prosecution

The use of the quant in the experimental phase was particularly problematic because the use of serial dilution can produce inaccurate results, particularly at the low end of the DNA amounts of tested. It is akin to the problems of splitting samples into replicates in low copy number testing.[74] The pool of available molecules within a solution is often simply too small at the lower quantities of DNA distribute evenly, or to create a serial dilution.

The population of fragments may be sparse and its distribution within the amplicon uneven. A criminalist does not pull out the correct proportionate amount of contributor DNA molecules with her micropipette when creating samples. The principle of multinomial distribution dictates that one cannot assume that if you began with 18pg and divided it into three replicates, there should be 6pg in each.

---

[74] The concept is straightforward in this elementary example:

> Given four marbles in a bag, two colored red, and two colored white, the odds that picking a red one at random will be are even: 1/2 . Not replacing the chosen marble to the bag alters the odds going forward, as the characteristics of the population of marbles in the bag has changed.
>
> If the first marble was red, the odds of the next one coming out red have diminished to 1/3. However, if the first marble was white, the odds of the next one coming out red are now substantially better: 2/3.

The likelihood of that outcome is only one in five.[75] Instead, you may have 12, or 7 or 3 in a tube.[76] The distribution will always vary. Thus, the drop-out rate will be different in each tube – and FST does not compensate this at all.[77] Because each aliquot will not have the same amount of DNA, the OCME practice of splitting samples into replicates causes FST to match runs to data that is inherently discordant.

In Mr. Johnson's case, where low template amounts of 119pg and 176pg are well below the ceiling at which current OCME protocols warn "caution" in interpretation of results[78], the effect is significant and observable. There is clear variation between each replicate test of the swabs: Certain alleles appear while others drop out; some peak heights are more robust, while others fade.[79]

Failing to account for this effect, the OCME goes a step further and assumes each replicate to be independent from each previous injection. Presuming the independence of this variable allows the OCME to multiply the replicates together, greatly empowering the numerator, and favoring the prosecutor's hypothesis. The use of "the product rule" here is the same math used to predict the probability of flipping a coin and having it come up "heads" five times in a row:

$$\frac{1}{2}\,x\frac{1}{2}\,x\frac{1}{2}x\frac{1}{2}\,x\frac{1}{2}\,=\frac{1}{32},\,or\;0.03125$$

But while each coin flip cannot not influence the next, such is not true with replicate testing of low template mixed samples, where each injection depletes the small population of PCR fragments in the vial of amplicon.

Accurately expressing the effect of sampling without replacement within an unordered population involves complex mathematics. For all such inquiries, there is difficulty in arriving at an unbiased estimation of the size of the remaining population.

---

[75] Budowle testimony, *Collins* Transcript, 12/9/2013, at 819-820.

[76] *Id.*

[77] *Id.*

[78] *See* OCME protocols.

[79] *See* electropherograms, attached at Exhibits G, H, I, and J.

		standard deviation of the estimator	usual estimator of the standard deviation of the estimator
μ	$\overline{X}$	$\sqrt{\frac{\sigma^2}{n}}$	$\sqrt{\frac{s^2}{n}}$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
p	$\hat{p}$	$\sqrt{\left(\frac{p(1-p)}{n}\right)}$	$\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$

Instead of accurately estimating these effects, the OCME ignored them. The defendant always bears the cost of these short-cuts, as they maximize their benefit to the H_p in the LR numerator. This over-estimates the weight of the evidence. Although perhaps consistent with the OCME's slapdash approach to FST, cutting corners in favor of the government is neither based on any reliable data nor any known scientific principle.

13. The validation study relied on an insufficient number of samples to estimate the drop-out rate

In OCME's FST article, they insinuated that thousands of samples had been used to calculate the drop-out rate. Presenting the drop-out study in the article, the authors report the constituent samples as aggregate amounts. (i.e. "In total, more than 2000 amplifications";[80] "In total, 350 mixtures and 104 touched items";[81] "In total, more than 557,000 test runs of the program were performed[.]")[82] Though perhaps impressive on their face, these grand totals are not the relevant numbers.

Considering the design of the drop-out study, which was predicated on the independence of drop-out rates for each comparison group, the overall number of amplifications is insignificant. Rather, it is the size of the comparison group that matters for statistical, population, and other research purposes. Starting with the 2,000 amplifications, that group was divided into at least 64 comparison groups. If the parameters identified in Mitchell's article are to be taken as the guide[83], testing

---

[80] *Id.* at 751.

[81] *Id.* at 752. When presenting the number of different categories of mixed samples, the sum total the authors present exceeds 350, which is not possible to reconcile from the text of the article. *Id.*

[82] *Id.* at 750.

[83] There is evidence in the validation studies to suggest that the number of amplifications reported in Mitchell's article was divided among a greater number of subgroups. Mitchell et al. For example, in Volume 10, the Identifiler 28 cycle runs were performed on a 3-contributor mixture set comprised of 3 proportions, not 2, as described in Mitchell's article. *Id.* In Volume 8, Identifiler 28 cycle runs were

was broken down between single-source and mixed samples. Single-source samples were then broken out into 8 testing groups by template size. Mixed samples were broken out into 2 and 3 contributor groups, each of which was further broken down once again into 2 ratio groups. Each of these resulting groups was then broken out into 6 testing group by template size. The entire array of single and mixed samples was broken out again by number of PCR cycles, one HT-DNA group subjected to 28 cycles, and the other LT-DNA group subjected to 31 cycles. The 2,000 amplifications in the study, divided by the 64 subgroups produces approximately 31 amplifications per group, a far smaller number per comparison set. This is not reported in the literature.

14. The validation study relied on an insufficient number of contributors to estimate the drop rates

Mitchell's article claimed that "samples of various combination of 85 contributors were amplified in duplicate or triplicate and analyzed for the purposes of drop-out and drop-in rate estimation."[84]

That statement is very misleading of the number of contributors they used to estimate their drop-out rates. As Dr. Mitchell admitted on cross-examination in the *Collins* hearing, the number was closer to ten.[85]

Alongside the claim regarding 85 contributors, the OCME also claimed those samples were racially diverse: "representing a mix of Caucasian, Asian, African American, and Hispanic ethnicities."[86] The inaccuracy of the number of contributors also calls the diversity claim into doubt. Moreover, the OCME has never put forward the racial make-up of the contributors.

It matters that OCME used only ten, rather than 85 or 100 contributors because of allele masking. Allele masking is exacerbated by the same combinations of contributors. The drop-out rate will therefore be distorted downward at certain loci in unpredictable random ways. The application of some correction like flattening

---

performed on a 2-contributor mixture set comprised of 7 proportions, not 2, as claimed. In both Volumes, the additional proportions are explained as being extant for other purposes in FST testing. Nevertheless, the key question is whether their amplification was counted against the total number of 2,000 amplifications claimed in the Mitchell article. See *id.* at 751.

[84] *Id.*

[85] Mitchell testimony, *Collins* Transcript, 5/1/2013, at 60.

[86] Mitchell et al. at 750.

out the drop-out rate or applying one standard deviation will not fix the inaccuracy of the data, because those corrections will not be applied in a uniform way.

15. Drop-in and other contamination are drastically underestimated

"Drop-in refers to alleles that are found in the mixture that were not present in the contributors."[87] This occurs when there is contamination and when there are artifacts from the amplification process ("stutter").[88] Again, the OCME's decision to use pristine samples in the lab, distinct from real-world samples, has decreased the reliability of the validation. Dr. Shapiro concludes that "Drop-in rates used by FST are also artificially low."[89]

Despite this, the OCME made the unsupportable and undocumented assertion in their FST paper that underestimating the drop-in rate is likely "conservative." However, if a defendant's profile shares many alleles with the evidence profile, a higher drop-in rate will lower the likelihood ratio.

The mock touch samples created for the FST validation displayed an enormous number of non-contributor alleles. For example, in the "clean touch" study (where a pen was cleaned with bleach, water, and alcohol to clear all DNA, then three known lab personnel touched it, and it was swabbed and tested), the mixture "Pen B" was tested.[90] First, there were alleles present that none of the known contributors had.[91] Second, PenB had a total of nine false positives, seven of which had LRs above both of the true contributors that had been defined as "cannot be excluded." One of those false positives, "JB" from John Butler's Caucasian American subpopulation of the NIST database, had an LR of 157 (i.e. 157 more likely to be JB and two unknown, unrelated individuals than if it were three unknown, unrelated individuals), which is rated as "strong evidence" on the OCME's qualitative scale.[92]

"Drop-in is much higher for "touch" DNA samples, FST under-estimates drop-in, drop-in creates false positives, and even a single drop-in can have very large effects

---

[87] Shapiro Report at 4.

[88] *Id.*

[89] *Id.*

[90] Shapiro Report at 5.

[91] *Id.*

[92] *Id.*

on LR."[93] This is another reason the FST's approach of empirically deriving drop-in rates from pristine samples is fundamentally flawed.

16. The OCME's degradation module did not work

The object of the drop study, to measure the rates of drop-out and drop-in, was foiled from the outset by a design fundamentally unsuited to the task. When this became apparent, the team tried to adjust, subjecting a module of samples to ultraviolet radiation to mimic real-world degradation effects. The results resembled evidence the lab works with every day. Testing the module however, showed FST didn't work. Presented with realistic drop-out, FST "did not increase the overall separation between true contributors and non-contributors …."[94] In other words, when accounting for a degraded sample's true level of drop-out, FST's LR calculation tended to falsely include non-contributors. Never putting the module into use, the OCME took the data from the degradation module and left the program unimproved.

The failure of the degradation module exemplifies the OCME's inexplicably cavalier and unscientific methodology in developing FST, and the bias that infects its core function. The LR calculations ought not to have failed when presented with degraded samples. To qualify for inclusion within the degradation module, samples had to meet benchmarks that are neither extreme nor uncommon.[95] All of the replicates at issue in Kevin Johnson's case, for example, meet these criteria.

Once the program did fail, however, the lab had the option to see the test as a successful experiment providing important information about how the model performs. False positive testing is an essential part of validation, and has never been seriously undertaken with FST. Although, after manipulating the data from the drop study, the OCME did run a number of samples against a database of non-contributors and logged a distribution of LR values that included a small number of false inclusions, this was never a true or rigorous false positive test. To understand how FST would perform with complex evidence samples, it would be necessary to subject it to extensive case-study analysis, locus-specific testing for

---

[93] *Id.*

[94] Mitchell et al. at 759.

[95] *See* Executive Summary at p. 5. Attached at Exhibit S.

the effect of different variables, and other rigors.[96] The degradation module test, while undertaken at too large a scale to qualify as sound methodology, still provided information.

The OCME wasted the opportunity. Later conceding that FST's ability to discriminate degraded samples was "an area in which further study is warranted[]"[97] may well be true, but also obscures other concerns. The unchanged variable in the degradation module experiment was not the degraded samples, but the LR calculator. That is what failed. Certainly FST cannot handle realistic drop-out. But the failure of the degradation module points to problems with biases embedded in the math. In a logarithmic calculation, each compounding bias, over-stating the evidence, restricting the H_d, and favoring the H_p accumulates. Those in FST were only revealed when the program was put under stress during the dropout module test, and run drop-out rates beyond expected limits. Despite such clear and significant implications – perhaps because of them – the team altogether abandoned the degradation module.

17. The absence of a case-specific false positive test is fundamental and alleged 'false positive' check was insufficient to bolster the credibility of FST's findings

"A 'false-positive' occurs when a LR above 1 is reported for a non-contributor's profile."[98] Again, the OCME opted not to use data from any given case sample to use in the testing of that sample. Instead, the OCME attempted to engage in a 'false positive study' during the validation stage. This is frowned upon by experts in the field. "It is impossible to accurately estimate the strength of a likelihood ratio without knowing the 'false-positive' rate associated with the sample."[99] Drs. Gill

---

[96] *See* Hinda Haned et al. *Validation of probabilistic genotyping software for use in forensic DNA casework: Definitions and illustrations*, 56 Sci. & Just. 104-108 (2016). Attached at Exhibit U.

[97] Mitchell et al. at 759.

[98] Shapiro Report at 2.

[99] *Id.*

and Haned have echoed this argument: false positive performance tests are necessary for responsible casework and exploration of factors going into the LR should be case-specific.[100]

The 'false positive study' or 'performance check' that OCME engaged in during the validation study was also flawed. "Due to design flaws in the 'false-positive' performance checks, the rate of false-positives produced by FST is unknown."[101] Dr. Shapiro details how the unknown characteristics of the contributors used in the study demonstrates its reliability.[102] False positives are far more likely – "sometimes drastically" – when individuals of a particular sub-population are used ("because individual sub-populations tend to share specific alleles and genotypes").[103] In contrast, if the sample was compared to a database that had contributors from different sub-populations than the sample, "then the false positive rate would be unrealistically lower."[104] Again, the OCME could reveal this answer by sharing their data – yet "the OCME never calculated the allele frequencies of this population, and refused to provide the racial make-up of the population. This information is critical in evaluating the validity of the false positive study."[105] Additionally, the number of contributors raises questions regarding the veracity of the false positive study. "Rather than use 1,252 different unrelated contributors, I found that over 70% of the contribution came from only 14 contributors, of unknown ethnicity (Figure 9). The extremely limited number of contributors to the FST validation means that not all the alleles in the general population are represented in these mixtures, and automatically excludes people even before a LR is calculated. This artificially biases the experiment against finding fortuitous inclusions (false positives)."[106] The OCME's claim that they conducted over 400,000 comparisons is misleading because the profiles of the

---

[100] *See* P. Gill & H. Haned.

[101] Shapiro Report at 2.

[102] *Id.* at 6-7.

[103] *Id.* at 6.

[104] *Id.*

[105] *Id.*

[106] *Id.* at 7.

same individuals were used "over and over again" – and therefore, the comparisons "were merely the same comparison over and over."[107]

At the *Collins* hearing, Dr. Rosenberg, a population geneticist and statistician at Stanford University during that hearing, testified that he carried out a number of simulations to analyze the FST's false positive study and determined that there are systemic variables affecting the false positive study because the number of false positives per bulk run does not match what you would expect if the false positives were randomly distributed.[108] Like Dr. Shapiro, he highlighted OCME's failure to investigate the role of population origin in the FST false positive study. He cautioned that the OCME should have analyzed the variables involved in the FST calculation to determine how each one affects the final result by holding all the variables constant and only changed one parameter at a time, to fully understand the importance of each variable.[109] Dr. Rosenberg concluded that his analysis showed that "we don't know enough about the false positive rate that would be relevant to a real world scenario."[110]

18. The OCME used varying injection times and machine sensitivity

The OCME performed its FST experiments on different machines, sometimes changing sensitivities and conditions in the DNA typing process.[111] These procedures also propagated an unknown amount of error in the calculation of drop-out and drop-in rates, which the OCME did not quantify in any demonstrable way.

19. The OCME failed to report simulated alleles used during validation

A glaring omission from the validation study and the Mitchell article is that the study authors had not truly amplified 15-locus profiles for each of the samples in the study, as they claimed. Rather, they had amplified a number of profiles used in the study only to 13 loci, a standard previously in wide use. To incorporate the additional loci, D2 and D19, into the study, rather than re-amplify the samples,

---

[107] *Id.*

[108] Rosenberg testimony, *Collins* Transcript, 12/5/13, at 67-68.

[109] *Id.* at 89.

[110] *Id.* at 90.

[111] Budowle testimony, *Collins* transcript, 12/9/2013, at 870.

the authors simply "simulated genotypes for these two additional loci and included them in the multi-locus DNA profile date in their validation studies."[112] Thus, an unknown number of profiles used in the study do not represent actual people. This obviously impacts the validity of any inferences that could ever be drawn from the study as to population statistics. According to Dr. Chakraborty, who was a professor in the Department of Molecular and Medical Genetics and Director of the Center for Computational Genomics at the Institute of Applied Genetics at the University of North Texas Health Science, at the time he wrote the attached affidavit, and was previously a member of the DNA Subcommittee, this hybridization is "statistically unacceptable because it produces pseudo-independence of genotypes of some loci in [the] data."[113]

The authors matched the boldness of this choice with the subsequent, greater omission: They did not disclose this fact to the Subcommittee. "[The] OCME did not reveal this approach during its presentations to the DNA Subcommittee. … If this fact was known during my participation on the DNA Subcommittee, my decision could have been drastically affected."[114]

20. Population effects ignored

The OCME's failure to recognize the well-established effect of population substructure is fundamentally unjustifiable. From some of the earliest applications like blood typing, the question of rarity has been forensic science's sibling inquiry. Statistically weighting evidence by reference to a population database raises the question whether that database is fairly representative of the population it is being for which it claims. Many issues, including sampling error, the size of the database, and the relatedness of those within it all affect definitions of frequency and can dramatically skew statistical calculations.

The OCME made three baseless assumptions regarding racial and ethnic substructure. Principally, against the evidence, the lab concluded that the population could be meaningfully divided into four racial categories, (three defined by region, one by skin color): Asian, Black, Caucasian, and Hispanic. These groupings are

---

[112] Declaration of Dr. Ranajit Chakraborty, filed in *United States v. Rashawn Smalls*, 14 Cr. 414 (BMC) (E.D.N.Y. Jan 22, 2015), Dkt. No. 29 ("Chakraborty Decl.") at ¶41. Attached at Exhibit O.

[113] *Id.*

[114] *Id.*

said to "represent the population of New York City." Where a New Yorker of Israeli, Pakistani, Indonesian or Moroccan-Arab descent would be categorized is not indicated.

Secondly, within each of these un-validated population tables, the OCME combines a jumble of unreliable self-reported data, skewed morgue samples, artificial numbers supplied by private companies, and computer-generated bin values. None of these databases actually sampled from any population at all. That the OCME could have confidence in its racial databases is belied by the lab's cavalier reporting of LR values that have no correspondence to the known race of the suspect. Neither "conservative" nor adequately corrective for FST's embedded unreliability, this counterfactual reporting – starkly at odds with every other judgment OCME weights against the defendant – raises more questions than it answers.

Third, the lab ignored substructure altogether in the design of the drop study. Had FST been designed to assess the weight of DNA results with accuracy, the OCME might have categorized each contributor in the drop study according to one of its four racial databases. With such plainly inadequate categories for measuring population substructure, it is understandable why the lab might decline to impose such definitions on volunteers. The failure to control for this data leaves FST practically inapplicable to real-world scenarios. For example, the program's unyielding requirement for every LR hypothesis FST calculates, whether H_p or H_d, that each contributor the mixture must share the same race. Such blunt tools do not and cannot assist the trier of fact.

21. Misrepresenting data, race, and the false positive study

The population geneticists that testified in the *Collins* hearing have identified significant problems with the FST's false positive test and have concluded that the design of this test is not generally accepted by the relevant scientific community.

Dr. Rohlfs, a population geneticist and post-doctoral candidate at University of California at Berkeley during the *Collins* hearing, and Dr. Rosenberg testified about the significant concerns they had with the FST. Both experts' opinions were that there were numerous instances where the OCME failed to investigate areas of the validation that led to unexpected results and that the OCME did not adequately consider race in their false positive test.

Drs. Rohlfs and Rosenberg focused their analyses on the false positive study because as population geneticists and statisticians their primary concern was to

determine whether the false positive study could accurately inform the relevant scientific community on how the FST would work in the real world. As part of their analyses they needed to know the race of the individuals that were used in the false positive study. Unfortunately, the OCME did not preserve the races of the contributors. This is scientifically indefensible given how important of a role race plays in forensic DNA testing. Dr. Chakraborty highlighted this issue in his testimony in *Collins*: That the racial background of contributors to mixtures is important to consider in validation, and that "[i]f race and ethnicity is not accounted" the results would not be acceptable.[115]

There are strong indications that the number of contributors in what the lab claims was the false positive study was significantly misrepresented, and drastically underrepresents minority populations. As Mr. Johnson is an African-American man, this issue of important concern.

22. Opaque reporting

Throughout the study, and beyond, the OCME has chosen to obscure rather than to clarify FST's calculations. There is nothing about LR calculations that requires secrecy or prevents full disclosure. In creating FST, the design team had every available tool to define how reporting would be crafted. The decision to report LR results in the aggregate – to not show drop-out, drop-in, and allele frequencies on a locus by locus basis – was a conscious choice.

Shrouded with mystifying scientific notation, but offering no practical information whatsoever by which counsel, the court, an accused, or another scientist can possibly reconstruct its results, an FST report spells intractable difficulty ahead. The only state court to fully appreciate the import of FST excluded it in *Collins*, but a combination of inertia and scientific intimidation allowed the OCME momentum with the program.

The Court's source code subpoena put an end to this.

VIII. Legal Standard: Fed.R.Evid. 702 and the *Daubert* Standard

1. The courts' gatekeeping function ensures that unreliable scientific opinion testimony does not poison the jury, by requiring that the proffered testimony be reliable and "scientifically valid"

[115] *See* Chakraborty testimony, *Collins* Transcript, 12/17/13, at 1119-21.

Under *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 596 (1993) and Rule 702 of the Federal Rules of Evidence, a district court functions as the gatekeeper for expert testimony. This "gatekeeping" role is to ensure that the jury's fact-finding mission is not poisoned by the introduction of unreliable scientific opinion testimony.[116] (Because "[e]xpert evidence can be both powerful and quite misleading because of the difficulty in evaluating it,"[117] courts must weigh the prejudice of expert testimony against probative force under Rule 403).[118]

Thus, as the *Daubert* Court explained, under Rule 702 and Rule 104(a), of the Federal Rules of Evidence, "the trial judge must ensure that any and all scientific testimony or evidence admitted is not only relevant, but reliable."[119]

Before proffered scientific evidence may be admitted, a court must make a preliminary assessment of whether it "is scientifically valid."[120] The Supreme Court in *Daubert* "made clear that, before scientific evidence can be said to 'help' under Rule 702, the connection between that the evidence and the facts in issue must be shown to be scientifically valid."[121]

"Rule 702's 'helpfulness' standard requires a valid scientific connection to the pertinent inquiry as a precondition to admissibility."[122] Accordingly, the trial judge's responsibility includes screening the reliability and relevance of expert testimony. First, the reasoning or methodology on which the proffered testimony is based must be scientifically valid. This requires a showing that (1) the underlying methodology and reasoning is grounded in "scientific knowledge," (2) that the

---

[116] *Daubert*, 509 U.S. at 595.

[117] *Id.*

[118] *See also Rink v. Cheminova, Inc*, 400 F.3d 1286, 1291 (11th Cir. 2005) (District courts are charged with a gatekeeping function "to ensure that speculative, unreliable expert testimony does not reach the jury under the mantle of reliability that accompanies the appellation expert testimony").

[119] *Daubert*, 509 U.S. at 589.

[120] *Id.* at 592-93.

[121] 29 C. Wright & V. Gold, FEDERAL PRACTICE AND PROCEDURE: EVIDENCE § 6263 (2d ed. 2016) (footnotes omitted). For example, the study of the phases of the moon "may provide valid scientific 'knowledge' about whether a certain night was dark," but "evidence that the moon was full on a certain night will not assist the trier of fact in determining whether an individual was unusually likely to have behaved irrationally on that night."

[122] *Daubert*, 509 U.S. at 591-92.

underlying findings were "derived by the scientific method" and (3) that the supporting work product itself amounts to "good science."[123]

Moreover, even if the testimony is indicative of scientific knowledge, it must be a scientific knowledge that can "properly be applied to the facts in issue" in the particular case.[124] The proffered testimony must have a "valid scientific connection to the pertinent inquiry as a precondition of admissibility."[125]

2. The court's gatekeeping function is to ensure that expert testimony is not presented to a jury unless it is reliable and relevant

"The main purpose of *Daubert* exclusion is to protect juries from being swayed by dubious scientific testimony."[126] To prevent the jury's fact-finding mission from being poisoned by the introduction of unreliable science, a court's "preliminary assessment" should probe the methodology and reliability of the bases of the expert testimony.[127]

"Faced with a proffer of expert scientific testimony, then, the trial judge must determine at the outset, pursuant to Rule 104(a) whether the expert is proposing to testify to (1) scientific knowledge that (2) will assist the trier of fact to understand or determine a fact in issue."[128] Under Rule 104(a): "Preliminary questions concerning the qualification of a person to be a witness, the existence of a privilege, or the admissibility of evidence shall be determined by the court, subject to the provisions of subdivision (b) [pertaining to conditional admissions]. In making its determination it is not bound by the rules of evidence except those with respect to privileges."[129]

In this preliminary determination, "proffered expert testimony should be excluded if it is speculative or conjectural; the [a]dmission of expert testimony based

---

[123] *Id.* at 590, 593-94.

[124] *Id.* at 592–93.

[125] *Id.* at 591–92.

[126] *In re Zurn Pex Plumbing Products Liability Litigation*, 644 F.3d 604, 613 (8th Cir. 2011); *see also Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 147, 151-52 (1999) (a court's gatekeeping obligation also applies to non-scientific expert testimony).

[127] *Daubert*, 509 U.S. at 592-93.

[128] *Id.* at 392 (footnote omitted).

[129] Fed.R.Evid. 104(a).

on speculative assumptions is an abuse of discretion."[130] As the Advisory Committee Notes accompanying Rule 702 state: "[t]he more subjective and controversial the expert's inquiry, the more likely the testimony should be excluded as unreliable."[131]

3. Rule 702 requires more than what is required by the relevance rules

One determination that *Daubert* requires is a relevance determination – i.e. that the proffered evidence would "assist the trier of fact to understand the evidence or determine a fact in issue".[132] This inquiry goes to whether there is a valid scientific connection between proffered testimony and a pertinent inquiry in the case. This requirement is otherwise known as the "fit" test: the proffered expert testimony must be sufficiently tied to the particular facts at issue in a case to be of aid to the jury.[133]

Although this is a relevance requirement, "Rule 702 demands *more* than what is required by the relevance rules" because the proffered expert testimony also "must be shown to be scientifically valid."[134] "This means that merely establishing Rule 401's broad 'any tendency' connection between scientific evidence and the fact it is offered to prove is insufficient to justify admission."[135] Instead, "[a] showing must be made that the evidence has the earmarks of valid science."[136]

4. The text of Federal Rule of Evidence 702

---

[130] *Major League Baseball Props., Inc. v. Salvino, Inc.,* 542 F.3d 290, 311 (2d Cir. 2008) (citation and internal quotation marks omitted); *Boucher v. U.S. Suzuki Motor Corp.,* 73 F.3d 18, 21 (2d Cir. 1996) ("expert testimony should be excluded if it is speculative or conjectural, or if it is based on assumptions that are so unrealistic and contradictory as to suggest bad faith or to be in essence an apples and oranges comparison") (citations and internal quotations omitted).

[131] See Advisory Committee Notes accompanying Fed.R.Evid. 702 (2000 Amendment).

[132] *Daubert*, 509 U.S at 590-91.

[133] *Id.* at 590-92.

[134] 29 C. WRIGHT & V. GOLD, FEDERAL PRACTICE AND PROCEDURE: EVIDENCE § 6263, at p. 194 (1st ed. 1997) (footnotes omitted).

[135] *Daubert*, 509 U.S at 195.

[136] *Id.*

Rule 702 provides as follows: A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:

> (a) the expert's *scientific, technical, or other specialized knowledge* will help the trier of fact to understand the evidence or to determine a fact in issue;
>
> (b) the testimony is based on *sufficient facts or data*;
>
> (c) the testimony is the *product of reliable principles* and methods; and
>
> (d) the expert has *reliably applied* the principles and methods to the facts of the case."

Fed.R.Evid. 702.

Thus, even if a witness qualifies as an expert whose testimony would assist the jury under subdivision (a) of Rule 702, "three subdivisions of Rule 702 impose on the trial judge additional responsibility to determine whether that testimony is likely to promote accurate fact-finding."[137]

These subdivisions – (b), (c), and (d) of Rule 702 – explicitly condition the admission of expert testimony on a showing that the expert's testimony is based on "sufficient facts or data;" that it is the "product of reliable principles and methods;" and that the expert has "reliably applied" those principles and methods to the facts of the case."[138] These subdivisions were added to Rule 702 post-*Daubert* in 2000 to incorporate *Daubert's* higher standards of admissibility by requiring that expert testimony be grounded on sufficient facts and data and be "the product of reliable principles and methods" that was "reliably applied … to the facts of the case."[139]

---

[137] 29 C. WRIGHT & V. GOLD, FEDERAL PRACTICE AND PROCEDURE: EVIDENCE § 6266, p. 302 (2d ed. 2016) (quoted parenthetically in *Whole Woman's Health v. Hellerstedt*, —— U.S. ——, 136 S.Ct. 2292, 2317 (2016)).

[138] Fed.R.Evid. 702 (b)-(d).

[139] When *Daubert* was decided – in 1993 – Rule 702 required only that an expert possess scientific or other specialized knowledge sufficient to assist the trier of fact to either understand the evidence or determine a fact in issue. The old Rule 702 read as follows:

5. "Law lags science; it does not lead it"

Importantly, the *Daubert* Court, in concluding its opinion, noted that while open debate is an essential part of both legal and scientific analyses, "there are important differences between the quest for truth in the courtroom and the quest for truth in the laboratory."[140] "Scientific conclusions are subject to perpetual revision. Law, on the other hand, must resolve disputes finally and quickly."[141] The scientific project is advanced by the wide-ranging consideration of a multitude of hypotheses, since "those that are incorrect will eventually be shown to be so, and that in itself is an advance."[142] But "[c]onjectures that are probably wrong are of little use, however, in the project of reaching a quick, final, and binding legal judgment—often of great consequence—about a particular set of events in the past."[143]

The balance struck by the Federal Rules of Evidence, therefore, is "designed not for the exhaustive search for cosmic understanding but for the particularized resolution of legal disputes."[144] As Judge Posner observed, "the courtroom is not the place for scientific guesswork, even of the inspired sort. Law lags science; it does not lead it."[145]

6. The Court must specifically find that the witness is an expert and his or her proffered testimony is also admissible

---

> If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise.

Daubert, 509 U.S. at 588.

In 2000, the Rule was amended to incorporate explicitly *Daubert's* higher standards of admissibility. 29 C. WRIGHT & V. GOLD, FEDERAL PRACTICE AND PROCEDURE: EVIDENCE § 6263 (2016) (quoting Rule 702 (c), (d)); see 28 U.S.C. Rule 702 (Pub.L. 93-595, § 1, Jan. 2, 1975, 88 Stat. 1937; Apr. 17, 2000, eff. Dec.1, 2000; Apr. 26, 2011, eff. Dec. 1, 2011).

[140] *Daubert*, 509 U.S. at 596-97.

[141] *Id.*

[142] *Id.*

[143] *Id.*

[144] *Id.*

[145] *Rosen v. Ciba-Geigy Corp.*, 78 F.3d 316, 319 (7th Cir. 1996)

Before proffered scientific evidence is placed before the jury, a court must make several determinations. First, the witness must actually be "qualified as an expert."[146] Second, the substance of the proffered testimony must amount to "scientific knowledge."[147] Third, even if the testimony concerns valid science, the reasoning and methodology underlying the science must be applicable "to the facts in issue" in the particular case.[148] And fourth, if the potential for the proffered testimony to mislead the jury and usurp its fact-finding function, outweighs its probative value, it should be excluded under Fed.R.Evid. 403.[149]

IX. FST results fail to meet standard for admissibility under *Daubert* and the Federal Rules of Evidence

1. FST results are not relevant

Can the proponent of evidence also offer a statistical proof that falsely claims to evaluate the entire universe of possible scenarios with the evidence, while also endorsing their own testimony with a qualitative grade? If this were remotely proper, witnesses would be endorsing themselves every day. It is obviously improper.

This, however, is precisely what FST purports to do. The program fails to produce relevant results because it embraces too much. By design, FST over-reaches to invade the jury's domain. In its framing of the LR, it misstates the law applicable to a criminal trial, and shifts constitutional burdens.

Yet, simultaneously, FST accomplishes far too little. To "explain" evidence samples taken from swabs of the gun here, FST assumes that drop-out occurred where genetic markers were not detected. Dispute persists whether drop-out should ever be assumed to explain an evidence sample.[150] Still, FST *always* assumes allelic drop-out: The program is custom-designed to explain evidence in these terms. Yet it does so inaccurately.[151] Drawing too close a focus on one possible variable – drop-out – FST under-values or misses other important issues altogether.

---

[146] Fed.R.Evid. 702.

[147] *Daubert*, 509 U.S. at 592-93.

[148] *Id.*

[149] *Id.* at 595.

[150] *See* Declaration of Dr. Dan Krane, October 30, 2016 ("Krane Decl."). Attached as Exhibit B.

[151] *See* Mitchell et al.

Whether assuming too much, too little, or entirely ignoring crucial factors, the program undervalues critical issues like race and ethnicity, family and relatedness, the frequency of observed alleles, the number and identity of alleles that may potentially be shared among contributors, and the effect of error the process.

Each of these variables affects the strength of the evidence. Put bluntly, they favor the H_d. The OCME's design of FST to not account for them, its testing and putative validation of the program without revisiting the formula in light of challenging data, and its implementation and use of the program here defy sound methodology. The program is not suited for the use to which it is being put.

2. FST's LR is not helpful in any way

Excising such plainly consequential information from its calculus, FST cannot produce results that have "any tendency to make a fact more or less probable than it would be without" its introduction.[152]

FST does not calculate a LR that truly explains the evidence. It does not attempt to do so. It is a proprietary means of organizing and applying a fixed set of data to a fixed formula. This is not "probabilistic genotyping." Its exclusive aim is to characterize whatever evidence an examiner uploads into it in light of the OCME's 2010 drop-out study. Given the degraded condition of the samples in Mr. Johnson's case, however, that study is of no utility here. Nor is the likelihood ratio FST produces.

While it would unjustifiable to conduct any science with the methodology the OCME employed with the drop study, the fact that FST is utterly unique in the world makes its sound validation all the more important. Validation requires more than running a program to see if it can generate a number. Particularly for the hypothetical calculations undertaken in a LR which, by their nature are not objectively verifiable, sound validation is a matter of methodology and technique that must proceed carefully and in stages. Studies have found several key steps to building a reliable LR-generating tool, none of which appear to have been followed here:

(1) Preliminarily validate the conceptual model and formalize the mathematics. This step is particularly amenable to peer review through publication.

---

[152] Fed.R.Evid. 401(a).

(2) Identify the theories and assumptions underlying the model and the mathematical and test each one for accuracy. The goal is not simply to test for success, but to test for failure; to test across cases and to alter variables within a single case study to perceive the robustness of the LR formulation, with testing for single-locus LR values preferable to full-profile LRs as the effects are easier to perceive; benchmarking and other comparisons with alternative methods of formulating LR values is critical to seek convergence towards similar results, although between platforms, differences in log values of an order of magnitude are considered negligible.[153]

(3) Faithfully implement the model in software, verifying that the program is running as expected through locus-level and controlled-variable tests. Appropriate evaluation of the proposed methods should consist of studies on a wide range of mixtures exhibiting different properties conducted by multiple groups *not associated with the developers*.[154]

The "empirical" drop study is among FST's defining characteristics: "Our approach is unique in that dropout and drop-in rates have been empirically estimated using our laboratory's casework protocols."[155] The OCME followed none of these steps. Rather, its validation study consisted of a paper-thin effort to ensure that the program would be implemented for casework.

X. Source Code

But obscured until the source code was ordered reviewed was just how deceptive was the entirety of FST's putative validation, its supposed review before the New York State Forensic Science DNA Subcommittee, the two articles by its authors and designers, and the whole of two admissibility hearings pursuant to *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923). In this record, as well as that before this Court, countless representations had been made as to FST's transparency, its reproducibility, and its lack of bias. In opposition to Mr. Johnson's application to inspect the code, the Government argued: "This is their program and they shouldn't have to disclose it just because an expert who is trying to challenge it

---

[153] *See* Haned. (Although her publication does not support her past views, Dr. Haned testified for the prosecution in *Collins*.).

[154] *See* PCAST report at p. 79.

[155] *See* Mitchell et al. at 756.

says I don't want to do the math by hand myself, because that's basically what it is."[156]

Our review of the code proved precisely the opposite. Mr. Johnson's experts could never have 'done the math by hand,' because the program completes unexpected functions about which the validation study is silent and the results of which are never reported to the subject of the testing.

1. **Undisclosed behaviors: discovered FST code function to throw out data at loci where [redacted] of the allele frequencies are observed**

Now that an initial review of the FST source code has been conducted, it has come to light that the program in fact performs LR calculations subject to a formula that has never been reported, and which favors the prosecutor's hypothesis.[157] Explained below, the embedded code jettisons data for entire loci when the frequency for individual alleles is considered to be unacceptably *high*. A high frequency allele would push a LR closer to 1. As FST is supposed to calculate the potential frequencies for every possible genetic combination at each locus, eliminating a locus entirely where data includes a high frequency allele is likely to favor the government. Following the data reported in the validation study and *Mitchell et al.*, eliminating an entire locus with a high-frequency allelic combination will favor the government in every instance.

This powerful, highly influential feature of FST has never been reported anywhere and has only been revealed in the review of the source code. It is axiomatic that it has never been subjected to even minimal peer review. To the contrary, indications are that strenuous effort has been expended to prevent it.

Nowhere in any of the FST documentation – the testimony to the DNA subcommittee, the published papers, the lab validation studies, the state court *Frye* hearing transcripts, all state and federal court trial testimony on FST, the affidavits from Dr. Mitchell or Dr. Craig O'Connor in federal cases, webinar presentations to the forensic community on FST, etc. – is there any indication that this function existed, or was part of the FST methodology.

---

[156] Tr. May 9, 2016, p. 42, ln. 2-5.

[157] Adams Decl. at section 5.6.

Likewise, there is no indication whatsoever that anyone at OCME even knew of the [redacted] data-discarding function, other than whoever coded it into FST.[158] In other words, the operator doesn't record this phenomenon – nor are they included in a subject's FST results. It is a function of the software that was previously undetected.

Like the other manipulations of FST that strayed from the observed data in their drop-out experiments, there is no record whatsoever that the [redacted] data-discarding function was tested to see if it might be prejudicial to a defendant. As Shapiro and Adams have shown us, variables should be case-specific and sometimes will hurt the defendant.

Disregarding loci with combined allele frequencies of [redacted] of the population is all the more startling given OCME's representations of the precision of their likelihood ratio. The OCME reports their LRs to three significant figures. The OCME's practice of reporting its likelihood ratios out to three significant digits implies a false sense of precision. This reporting gives the jury the idea that the method is very precise (which can easily be conflated with accuracy). The presentation of a statistic expressed to three significant figures lends undue weight to the reliability of its figure, especially since it is an imprecise value.

Dr. Caragine[159] claimed that the lab could measure drop-out on a locus by locus basis. Thus the OCME validation team decided that the locus-specific drop-out

---

[158] Indeed their publications affirmatively suggest otherwise. In their paper describing the method and validation efforts, it is stated that FST uses drop-out rate (minus one standard deviation) is used for each locus. *See* Mitchell et al. at 752 ("In order to be conservative, FST uses the drop-out rate estimate minus one standard deviation for each locus, template DNA quantity, number of contributors, and ratio for mixed samples.")

Moreover, Mitchell testified:

> Q: Would locus-by-locus reporting help troubleshoot any problems in the input of the values, for example, for the FST?
>
> A: No, because they're all reported, they're all part of the report. You can see what went in.

Mitchell testimony, *Collins* Transcript at 795, line 6-11.

[159] Dr. Caragine was the Deputy Director of the OCME's forensic biology department when FST was created. She resigned upon being confronted by two instances where she breached protocol by reassigning cases when she disagreed with findings. This behavior lead her to be the subject of an investigation of the State of New York Office of the Inspector General in December 2013.

rate could be determined empirically by running samples in the lab and counting how many times they saw drop-out.[160] According to its developers, the quantitation-based, locus-specific drop-out rate made FST unique in the world.[161] In fact, in their papers and in their testimony in state court *Frye* hearings, Drs. Caragine and Mitchell compared the FST favorably to the previous practice at the lab of assigning a combined probability of inclusion to complex DNA mixtures where drop-out was suspected.[162] Drs. Caragine and Mitchell in their *Collins* testimony cited to scientific publications indicating that the practice of ignoring is potentially prejudicial to the defendant.[163]

They suggested that their approach was accurate because they had done so many tests in the lab to figure out the locus-specific drop-out rates.[164] But these numbers were deceptively inflated.

2. Flawed software engineering

Nathaniel Adams explains the importance of testing software "on its own, in accordance with strictly defined test criteria developed from its own requirements and specifications documents."[165] "Using IEEE's definition of verification, 'The process of evaluating a system or component to determine whether the products of a given development phase satisfy the conditions imposed at the start of that phase,' it is difficult to assess what verification processes have been undertaken during FST's validation and continued use."[166] No explicit protocols have been

---

[160] "We counted how often we saw drop-out in thousands of samples." Caragine testimony, *Collins* Transcript,12/13/2012, 29.

[161] *See id.* at 112-113 (discussing locus specific drop-out rates difference between FST and two other probabilistic genotyping programs).

[162] *Id.* at p. 78-79 ("The combined probability of inclusion cannot factor the drop-in and drop-out cannot accommodate the random match probability…if you have drop-out, it lowers the weight of the statistics, when we are talking about the comparison of known profile to a mixture that we have not pulled out the DNA profile of the contributor then we need a different statistic, the likelihood ratio it accounts for drop-out and drop-in.") ("The likelihood ratio allows us to put statistics on more types of comparisons, comparison where we see one allele if [sic] suspect's profile that is not seen in the mixture so now we can put a statistics [sic] on that where you could not do that with the CPI").

[163] *Id.*

[164] Mitchell et al. at 750; *see also* Mitchell testimony, *Collins* Transcript, 12/13/2012, at 126 *et passim*.

[165] Adams Decl. at 5.2.2.

[166] *Id.* at 5.3.

provided.[167] The FDA further recommends that software testing be performed independently i.e. "performed by an organization free from control by the supplier, developer, operator, or maintainer."[168]

The stylistic conventions of code do not follow one style guide.[169] No specific style guide was included in the validation study nor any other material that Mr. Adams was provided.[170] The risk in said absence is "Without explicit descriptions of intended behaviors, reviewers and developers are left to infer the purpose of the code from its own function. In the absence of external definitions of intended functionality, a developer runs the risk of justifying the behavior of existing code by its sheer existence."[171] Style guides help prevent "code smells" – "a *smell* is not a defect in itself but is a deviation from good coding practices, which can indicate underlying software defects."[172] Adams found smells throughout the FST code. Written into the code, he found the following as a name of the variable:

[redacted][173]

Adams also found instances in the FST code where the function of the code itself was, at best, mislabeled, and at worst, didn't work. For example, in one comparison dropdown list, options #3 and # 4 could only be activated if, instead, Adams selected #5 and #6. In other words, the program required choosing one option in name in order to perform another option's function.

The inconsistencies that Adams found between the validation study's description of the code's behaviors and the source code raises two possibilities: first, that the source code to FSTπ is not the same program that was validated by OCME. If that's the case, then the validation of FST is irrelevant to FSTπ. Scientific validation and approval of FSTπ is necessary before it may properly be admitted under *Daubert*.

---

[167] *Id.* at 5.3.1.

[168] *Id.* at 5.3.2.

[169] *Id.* at 5.4.3.1.

[170] *Id.*

[171] *Id.*

[172] *Id.* at 5.4.3.2.

[173] *Id.*

Second, it may be that FSTπ has always been FST, but the program simply never did what its proponents claimed. Given that no such disclosure was ever made in FST's documentation, to the New York State Forensic Science Commission who approved FST, in testimony or published papers by the OCME, it is not possible to answer this with certainty. It is clear, however, that the record of validation – already dubious – is broken by FSTπ. It wholly undermines any reliance upon FST's validation to establish admissibility. Regardless of FSTπ's origin, FST cannot pass the *Daubert* test.

XI. The testing in Kevin Johnson's case

After the guns were seized in April 2015, six swabs were collected (three from each gun); four had insufficient material and two swabs contained a mixture of DNA without known contributors. The OCME ran a comparison against Mr. Johnson.

1. Slide grip grooves and mag release

The first swab, slide grip grooves and mag release, "yielded approximately 119 picograms of DNA (as determined by the OCME)."[174] FST generated a likelihood ratio: 156 times more likely that the mixture contained Mr. Johnson and one unknown person than two unknown persons, when the Caucasian allele frequencies are used.[175]

Two amplifications were performed during the testing. Dr. Krane found that the observed differences between the two sets of results "constitute a low degree of reproducibility for this evidence sample." [176]

A total of thirty-seven unique alleles were observed in the two genotypes generated from the "slide grip grooves and mag release" sample, across the fifteen autosomal loci tested. Twenty-one (57%) of the total unique alleles observed were observed in both sets of results. Each genotype generated from the "slide grip grooves and mag release" sample contained eight unique allele calls not observed in the other genotype generated from the same sample.[177]

---

[174] Krane Decl. at 6.

[175] As explained above, the OCME divides the population into four racial categories; yet, it does not test a subject against the race-based population he or she comes from.

[176] *Id.*

[177] *Id.*

Dr. Krane points out that there is immediate evidence of drop out in the sample: no alleles whatsoever appear at the FGA locus. He also explains that while the OCME reports its results in a way that includes a reference to a specific number of persons in a mixture (e.g. subject + 2 unknown persons = at least a three-person mixture), there is no record of how they came to said conclusion. As discussed above, number of contributors can be an arbitrary determination – fallible to human error. This determination is particularly important regarding mixtures with more contributors (e.g. a purported three-contributor mixture): the FST is not validated for 4-person mixture.

2. Front/back strap and grips #2

The second testable swab came from the front/back strap and grips #2. It "yielded approximately 175 picograms of DNA (as determined by the OCME)."[178] FST generated a likelihood ratio: 66 million times more likely that the mixture contained Mr. Johnson and two unknown persons than three unknown persons, when the Caucasian allele frequencies are used. Two amplifications were performed here. Again, drop out was immediately apparent because no alleles were present at the FGA locus. Dr. Krane again expressed the low degree of reproducibility for this sample, given the discrepancies between the two tests:

> A total of fifty-one unique alleles were observed in the two genotypes generated from the "slide grip grooves and mag release" sample, across the fifteen autosomal loci tested. Thirty (59%) of the total unique alleles observed were observed in both sets of results. Each genotype generated from the "front/back strap and grips #2" sample contained either ten or eleven allele calls not observed in the other genotype generated from the same sample.[179]

XII. As applied challenge: OCME's treatment of the DNA samples here was inconsistent with their own validation and protocols and is not admissible

The OCME's first step in this type of DNA analysis, as explained above, is the STR typing, using the Identifiler (which will also be retired in January, along with FST). Identifiler 28 (ID28) was used for the swabs in Mr. Johnson's case, which

---

[178] *Id.* at 7.

[179] *Id.*

means the DNA is amplified 28 times before testing.[180] The validation study implemented those requirements when FST was tested:

> Samples amplified using Identifiler 28 cycles: …
>
> Unless otherwise indicated, samples were injected according to the following parameters:
>
> a) "ID28 high (IR)": samples 200 pg and below – 1 kV for 22 seconds
>
> b) "ID28 normal (I)": samples > 200 pg – 5 kV for 20 seconds.[181]

The OCME's published protocols for forensic STR analysis set out clear directives for its use:

> For ID31, samples with less than 20 pg amped may be injected high immediately to reduce the number of reruns necessary.
>
> For ID28, samples with less than 200 pg amped may be injected at rerun parameters immediately as well.[182]

Amplification Cycle	Specification	Module Code	Parameters[183]
Identifiler 28	Normal	I	1 kV for 22 sec
	High	IR	5 kV for 20 sec

Here, the template amounts of the DNA on both swabs that were suitable for comparison were below 200 picograms: 175 and 119.

According to the validation study, such samples were validated using High ("IR") parameters. Indeed, in the 'Methods' section of the summary of Volume 10 of the Validation study, they used IR for 150 pg. samples.[184] So too was IR used in the validation study's drop-in and drop-out studies.[185]

---

[180] The alternative, Identifiler 31, which amplifies a sample 31 times, is used on low-template DNA.

[181] Executive Summary at p. 6.

[182] OCME Protocols at p. 160.

[183] Excerpt from Table 5, *Id.*

[184] Krane Decl. at 8.

[185] *Id.*

The operator in Kevin Johnson's case did not follow the testing protocols that were validated in the validation study. Instead, he used the "normal" specification and used the Normal ("I") parameters. As a result, the initial amplification was run at a different voltage, for a different amount of time than the samples in the validation study. Dr. Krane writes that "additional validation work" regarding injection parameters would be required to understand "the impact of this disparity on the accuracy or reliability of FST."[186]

Whether Dr. Krane can make a scientific conclusion about the import is not the question. At issue before this Court, under *Daubert* and 702 is whether the testimony sought to be introduced by the Government is the product of reliable principles and methods. If the methods used in Mr. Johnson's case do not rely on the methods that were actually tested, then the results at issue *in this case* must be excluded.

XIII. Conclusion

Any of one these issues should bar FST's admission, as the LR is uniquely open to bias and unfair prejudice. Combined, FST's record reflects a disregard of sound methodology so complete as to undermine even the relevance of FST's results. Their admission would violate the *Daubert* standard, as well as those of the Fifth and Sixth Amendments, and the Federal Rules of Evidence and Criminal Procedure.

For all of these reasons, and any other that the Court finds appropriate, we seek the exclusion of any and all evidence generated by, and testimony about, FST, or in the alternative, a *Daubert* hearing.

Respectfully submitted,

/s/
Christopher Flood, Esq.
Sylvie Levine, Esq.
Robert M. Baum, Esq.

Counsel for Kevin Johnson

[186] *Id.* at 9.